Thinking Machines and Electronic Publishing

Wide Area Information Servers Planning Meeting

February 26, 1992

Brewster Kahle
Ottavia Bassetti
Thinking Machines Corporation

"We do not do it because we want to.
We do not do it because it is permitted.
We do it because we are compelled."

Invitees: Sheryl Handler, Dick Clayton, Harvey Weiss, John Mucci, DaveWaltz, Jim Bailey, Bob Millstein, Ted Tabloski, Craig Stanfill, Brewster Kahle, Ottavia Bassetti

Executive Summary

1990 was for WAIS the year of the Peat Marwick, Apple, Dow Jones project: a way to build a sound know-how in the area of text management inside organizations.

1991 has been the year for developing new products (SEEKER-disk and SEEKER-timesharing) and the year of the Internet Experiment: a way to bring a CM to everyone that has access to the Internet (millions of people).

WAIS in 1991 has been the top hit in the electronic community. This has brought to Thinking Machines a lot of publicity, both in the electronic community environment and in many different markets (publishing, financial, government, academic).

WAIS is setting the way in the electronic publishing world and this reflects in a vast number of contacts which are turning into prospects for CMs and promising collaborations for Thinking Machines.

1992 holds many options in this area. They include consolidating the development effort by porting the text retrieval software to the CM5, continuing the Internet WAIS service, making WAIS a bundled application for CM users. We can decide how far to go in turning WAIS into a product to take advantage of our lead in electronic publishing.

Contents:

1990 PART 1
1991 Where We Stand:
Status and Achievements of the WAIS project

Text management in organizations:
the Peat Marwick project
WAIS, a supercomputer on every desk:
The Internet experiment
WAIS and TMC in the News.
CM software products and products for CM users
Collaborations for TMC
From many contacts, real prospects

1992 PART 2
Where We are Going:
Opportunities and Options for the WAIS project

Text Retrieval in Electronic Publishing
Text Retrieval in Data Management:
Why commit now? Continue to Lead in an Innovative Area as it Grows
Why commit now? Push to a Paradigm Shift in DBMS
Thinking Machine's Options for the Text Market
Level 1 OS Support for Text
Level 2 Text Application Prototyping
Level 3 Text Application Product Development
Level 4 Text Application Marketing

Where We Stand:
Status and Achievements of the WAIS Project

1.a. Text Management in Organizations:
The Peat Marwick project

The Peat Marwick project was a joint project between Apple, Dow Jones, Peat Marwick and Thinking Machines. The goal was to build an information system for Peat Marwick partners distributed in 6 different cities. It was started in April 1989, after 9 month it was installed and tested for a period of 6 months.

The system proved successful. The decision on the part of Peat Marwick not to go ahead and buy it depended on the high costs of networking the over 200 offices of the organization: an order of magnitude bigger than buying the biggest CM.

Two lessons were learned:

- WAIS is the way to go for an information system,
- Networks are not yet affordable in the US for highly distributed organizations.

1. b.WAIS, a Supercomputer on Every Desk:
the Internet Experiment

The Internet is one of the greatest opportunities existing on the US market in the race to the electronic infrastructure. It represents the main worldwide infrastructure for electronic communities. In the near future the Internet will go from being a government sponsored infrastructure to becoming a commercial endeavor.

WAIS was launched on the Internet starting April 1991. In only 10 months it has taken the lead and is setting the standard for what is going to happen in the electronic publishing world. This gives the project very interesting opportunities to get an early placement in the NREN projects and maybe even HPCC projects and in the rapidly changing and growing electronic publishing world.

Digitized by the Internet Archive
in 2026 with funding from
Brewster Kahle

https://archive.org/details/thinkingmachines00unse_7

For the Internet Experiment the following components were developed and made available:

- Interfaces for Macintosh, Xwindows, emacs
- Software for creating database servers
- A CM available 12 hours a day running a small set of databases
- A directory of servers

In 10 month, the following achievements :

1. Usage of the system: 10,000 people have used WAIS

- Over **10,000 people in 24 countries** have used WAIS
- **160 Databases** have been set up worldwide (US,Finland, Mexico, Netherlands, Norway, Canada, Singapore, etc)
- **100 Downloads/day** of software from think.com (not counting European and Japanese redistributors).
- **Applications** : - Campus information servers
  - Biological databases
  - Bulletin board indexing
  - Libraries

2. The community works for WAIS.

The reaction of the electronic community has been immediate and effort in contributing software and ideas has grown and does not seem to be slowing down.

Commercial interfaces under development:

- Apple
- NeXT
- Pandora

Freeware interfaces from other sources:

- dial-up interface (by NFS)
- NeXT interface (by U of Utah)
- MS-DOS interface (by UNC)
- MS-Windows interface (by UNC)
- Integration into AVS (a visualization package) (by NCSC)
- SunView interface (by UNC)

- VMS server and interface (by UNC)
- IBM-VM interface (by UNC)
- Prospero interface (by ISI)
- World Wide Web interfaces (by CERN)

3. WAIS is becoming a standard

The success of WAIS on the Internet and our work with the Z39.50 committee has set the way for WAIS to become a standard in the electronic publishing world. Two groups are promoting this effort:

Z39.50 Implementors Group (Standards Committee).

We have been pushing this standard to avoid the proprietary protocol period which would set the field back years. When we became part of this committee, in January 1989, there were only librarians on it. Now it is a major focus as the electronic publishing standard. We changed the standard, and the committee has adopted a version of each change, so to handle multi-media and large documents. Companies that WAIS has attracted and that are now active participants in the committee are: **DowJones, Mead Data, Dialog, Maxwell Online, Apple, NeXT, Sun;** in addition **10 major Universities, (UCLA, Berkeley, MIT, Carnegie Mellon, Harvard, etc.), Library of Congress, NLM, OCLC, Chemical Abstracts,** etc.
The committee meets every 4 months, and we are close to a version of the protocol we can use: Z39.50-1992.
Contacts:
Brewster Kahle (brewster@think.com)
Mark Hinnebusch (fclmth%nervm.bitnet@uga.cc.uga.edu)

Z39.50 Testbed as part of the Coalition of Networked Information.

A smaller group that just formed based on our insistence that a faster moving group meet regularly to get something done. The group is a subset of the Implementors group formed by commercial companies that want to adopt the protocol for mainly non-library uses. The goal is to move faster than a regular standard committee.
Contacts:
Paul Peters, Director CNI padler@umdc.umd.edu
Clifford Lynch, UCAL calur%uccmvsa.bitnet@uccvma.ucop.edu

4. WAIS Research and Support Center started in North Carolina

Kudzu: North Carolina WAIS Research and Support Center started spontaneously in North Carolina as part of MCNC. **The first meeting was on Feb 2-3 1992 and was sponsored by an NSF grant.** We have blessed their initiative while at the same time keeping control over the directions of the software development. Our goal is to have a center stimulate and coordinate the efforts going on in the electronic community to develop software for serial servers and clients. We have been very clear in our desire to place this center at a site where a CM exists and they are looking for funding at various levels.

This is a major achievement in the WAIS program. North Carolina MCNC center sees WAIS as a strategic technology for itself and is looking to help in assuring compatibility between vendors from a non-threatening non-commercial vantage point. They are serving this role with AVS, a visualization package, that runs on many mini-computers and super-computers (and maybe a CM soon).

Since MCNC has the best digital network in the country, and since this WAIS center is in the department that constructed it, we will get to use the best technology for WAIS experiments. **The first "Video WAIS" system** will be done there.

Contacts:
Brewster Kahle (brewster@think.com)
George Brett (ghb@concert.net)

1.c. WAIS and TMC in the news

WAIS has been easily and heavily covered by the press: it's an application easy to understand in its usefulness. Our strategy has been of openness and we think it has contributed positively in keeping up the image of Thinking Machines as an innovative, research oriented company.

Articles and papers in 1991:

NYTimes	"For the PC User, Vast Libraries" July 91
Byte Mag.	"Browsing through Terabytes," May 1991
MacWeek	"WAIS Promises Easy Text Retrieval", May 1991
Esther Dyson	"Client-Server Standards for Text," Rel 1.0 91
San Jose Mercury	"Network to Unite Data Bases," July 1991
UNIX Today!	"The Promise of the WAIS Protocol," Dec 91
Int'l Herald Trib	"Having it All: Vast Data Networks Near," Jul 91
Online Magazine	"An Information System for Corp Users: WAIS"
Electronic Networks	" An Exec Info Sys for Unstructured files: WAIS"
ASIS	"WAIS Interfaces", TMC, Apple, NSF. Feb 92
Apple Library	"Electronic Publishing and Corporate Librarians"
Digital Media (Seybold)	"Electronic Publishing and Public Libraries" 92

TV appearances in 1991:

CNN

1.d. CM software products and products for CM users

During 1991, we have developed a full product line of WAIS server technology for 100MBytes to 100Gbytes databases. This development has been done with very limited staff and in a short time since the product functionality was closely defined. This work was done in addition to the freeware releases which served as marketing and advertising.

CM2 WAIS servers:

• SEEKER-RAM-T
Timesharing RAM based software

Runnable on any CM2 system, this software runs the WAIS software in the background and answers questions while other applications are running. This is targetted to all CM2 sites as the first turn-key CM application. Adding new databases to the server is easy to do.

Status: Ready 2/92
Author: Tracy Shen in 6 man-months

• SEEKER-DISK

This system can serve very large databases in a cost effective manner by using the Datavault as an active component to the system. We found this product was needed through discussions with information providers and through the Peat Marwick project. For this product we have developed a boolean search algorithm.

Status: under test 2/92. Ready 4/92.
Author: Harry Morris in 18 man-months

For CM customers:

• WAIS interfaces and serial servers have been distributed, as unsupported software, in the latest release for CM2.

•PRISM has integrated the WAIS interface for CM documentation

1.e. Other Companies commit to WAIS

•**Apple** is spending $1 million a year on interface development that will be compatible with WAIS. Going to product on WAIS enabled interface (confidential). This is a follow-on from the 1990 phase of the WAIS project.
Contacts:
Brewster Kahle (brewster@think.com)
Thomas Erickson (thomas@apple.com)

•**NeXT** will integrate WAIS into their DB toolkit for every machine they ship starting with release 3.1. This is a significant event for Thinking Machines and WAIS and has been the result of a long relationship.
Contacts:
Brewster Kahle (brewster@think.com)
Adam Hertz (Adam_Hertz@next.com)

•**Dow Jones** will make WAIS (DowQuest) available over DowVision.This means that Dow Jones will make the DowQuest service available via the WAIS protocol over the DowVision network. DowVision is a 9.6kb lease line network for delivering a wire feed of data and allowing people to contact DowJones services that way rather than through a Public Data Network such as Telenet and Tymnet.
Contacts:
Brewster Kahle (brewster@think.com)
Greg Baber (609-520-4000)

•**Mead Data** is implementing their own Z39.50 product. Mead Data is the second largest information provider in the US (revenues around $224). They joined the Z39.50 committee because of us, but they are wedded to their hardware base. We have talked about them using our machines, but no signs of hope yet. We only now are starting to have a product that would make sense to a Mead Data, so it is not too surprising.
Contacts:
Brewster Kahle (brewster@think.com)
Peter Ryall

•**Telebase** is buying a machine (8k CM2) to run the "Homework Helper" service for K-12 children with newspaper, encyclopedia, cliff notes, atlases etc. This will be marketed over Compuserve and other networks starting Q1 93.
Contacts:
Ottavia Bassetti (ottavia@think.com)
Carol Bee Laty (carol@think.com)
Lawrence A Husick (law.husick@AppleLink.Apple.COM)

•**Pandora Systems.** Pandora is a west coast software development company that specializes in user interfaces for information services over phone lines. They have done a design for a WAIS interface and are looking for support for the development. They have submitted a grant proposal to NSF for k-12 work, and they are negotiating with Telebase to develop an interface for them.
Contacts:
Brewster Kahle (brewster@think.com)
Mark Graham (mark@pandora.sf.ca.us)
Mitra (mitra@pandora.sf.ca.us)

1. f. Collaborations for TMC

1. On the WAIS side

• **Sun Microcomputer.** Sun is implementing Z39.50 because of our suggestion to use the protocol. We are not guaranteed to be compatible, but we are trying somewhat. They have product plans to use this system for customer support and bug tracking. They are ahead of us in using this technology to help them internally and to serve their existing customers.
Contacts:
Andy Bensky (andy.bensky@corp.sun.com)
Brewster Kahle (brewster@think.com)

•**High Performance Computing and Communication (HPCC)** Software sharing project will be using the WAIS protocol for transmitting source code, images, data files, text, etc. This is a large project with lots of funding, and we are just starting on the collaboration. The project manager is up for directing some money towards us, but we have not figured out what to do yet.
Contacts:

Brewster Kahle (Brewster@think.com)
Barry Jacobs (bjacobs@gsfcmail.nasa.gov)

• **SystemHouse** is a software integrator that is working with Dow Jones to solve their internal information needs. They are proposing to put up WAIS on the NeXT machines for each of the reporters and editors. We would be helping them interface to the DowQuest system when the time came.
Contacts:
Brewster Kahle (brewster@think.com)
Tracy Shen (shen@think.com)
John Coyne SystemHouse 212-255-8322

• **Xerox: Aphrodite project.** A project to make smart servers that contact servers automatically and filter information based on extensive evidence of user preferences. They are planning to use WAIS as the protocol and search engines. If they get internal funding they would like to pay for one day a week of Jonathan Goldman's time to unite the two projects. The benefit would be that a major Xerox project would start using the protocol and that solid research on smart servers would be made and operated on the Internet to add value to the WAIS system. The details have not been worked out.
Contacts:
Jonathan Goldman (jonathan@think.com)
Jeff Shrager (shrager@parc.xerox.com)

• **Microsoft**. Their "Information at your finger tips" group is very interested in WAIS. I have not been very forthcoming since they have a reputation in Silicon Valley of stealing ideas. Eventually I will talk with them.
Contacts:
Brewster Kahle (brewster@think.com)
Edward Jung (edwardj@microsoft.com)

• **Stanford University - Terry Winograd**: Prof Winograd and his students are using WAIS as a basis for developing a high-level generic interface to the Internet. In particular they are interested in user interfaces for information retrieval. Their emphasis has been on adding structured document types and searches to the existing WAIS tools.
Contacts:

Terry Winograd (Winograd@cs.stanford.edu)
Harry Morris (morris@think.com)

•**Stanford University:** Bo Parker heads the **SPIRES database consortium** at Stanford. They are about to apply for a DARPA grant to serve computer science journals over the Internet. They plan to use abstracts and OCRed text for searching, and present a bitmap of the page image to the user. They have chosen WAIS as their underlying search and retrieval technology. This proposal may result in the sale of a CM, or a research contract for Thinking Machines. In either case, it will result in the creation of a high profile database, and exploration of delivery of bitmap images. There is also a good chance that Stanford researchers will contribute IR work as part of this project.
Contacts:
Bo Parker (bo_parker@Forsythe.stanford.edu)
Harry Morris (morris@think.com)

•**US Geological Survey** is now using WAIS to distribute information and maps internally. They have extended client programs to suit their needs. They have made a decision to use WAIS as an additional access to one of their production systems. This is very encouraging.
Contacts:
Brewster Kahle (brewster@think.com)
Tim Gauslin "[E.CHRISTIAN/OMNET]MAIL/USA%TELEMAIL"
@INTERMAIL.ISI.EDU

•**UNC Chapel Hill Campus Information System** is using WAIS as it's backbone. They have put up 24 WAIS servers of information about their campus as well as astronomy images, recipies, and customer support information. They have been very active in making new user interfaces for MS-Windows, VMS, and IBM-VM. Their work on WAIS has been invaluable.
Contacts:
Jim Fullton (fullton@rhumba.oit.unc.edu)
Brewster Kahle (brewster@think.com)

•**Johns Hopkins University**: new Mac interface and key biology databases The biologists at Johns Hopkins have made a key biology database available with WAIS, and are making a new Macintosh interface based on hyper-card. They are committing programmer resources to making WAIS useful to biologists.
Contacts:

Francois Schiettecatte (francois@hel.welch.jhu.edu)
Harry Morris (morris@think.com)

•**Finland Biology**: Rob Harper in Finland has put up 14 key biology database and has even gone so far as to make their own directory of servers. The biological use of WAIS has been the most promising on the internet in terms of a professional set using the existing tools to do their jobs. This work has introduced the Connection Machine to them, and can help them justify a machine that is under discussion now. The other key biology databases are at Intelligenetics, in Singapore, and on the Connection Machine (because of Rob Jones, our geneticist here).
Contacts:
Rob Harper (harper@nic.funet.fi)
Brewster Kahle (brewster@think.com)

•**Gopher, World Wide Web, Prospero, Archie**.There are several other Internet projects that overlap with WAIS; we are all working together. **Gopher** is a system for finding information and services on the internet from Minnesota. **World Wide Web** is a project from CERN in Switzerland for reaching out to remote sources for FTP'able files. **Prospero** is an extension to the Unix file commands that allows users to retrieve remote files and to reorganize them. **Archie** is a database of all publicly available files on the internet and is easy to use from anywhere. This is a very popular system out of Canada. They are forming a company to commercialize it.
Contacts:
Brewster Kahle (brewster@think.com)
WWW: Tim Berners-Lee (timbl@nxoc01.cern.ch)
Prospero: bcn@ISI.EDU (Clifford Neuman)
Archie: Alan Bajan (bajan@mocha.cc.mcgill.ca)

2. Collaborations on the CM side

•**SAIC Evaluation for CIA**. The CIA asked SAIC to evaluate our search engine. They are working with MRJ in doing this.
Contacts:
Duane Dregits SAIC
Barbara Lincoln (barbara@think.com)

•**Baylor College of Medicine** is one of the 4 centers that are on a long term NSF grant to create an environment and an information service for the biomedical community. They are also working very

closely with the National Library of Medicine.They have integrated WAIS in their "Virtual notebook" environment and are using the CM2 rice machine to work on text retrieval tuning. They have already submitted proposals to NSF to get funding for CM research and would be ideal academic partners to work on information services for the medical community.

Contacts :
Tony Gorry (tony@wilkins.iaims.bcm.tmc.edu)
Ottavia Bassetti (ottavia @ think.com)

• **Telescope project North Carolina**: From Jim Fullton:"UNC just got $10,000,000.00 to build a big telescope in Chile, and we are looking for NASA/NSF money to develop a big, searchable network oriented database for image access. That should explain my interest in making image transport work."
"We are quite interested in using a CM, if we can get our funding together (which is by no means a sure thing). We want to merge images and text to create a really nifty research tool. We have been considering using our Convex, mainly because the big tape units work with it, but obviously WAIS is the big software selling point for the CM, as well as the fact that it will make a *much* better search engine. We just have to make sure that all the hardware will work together."

Contacts:
Jim Fullton (Fullton@mdewey.ga.unc.edu)
Brewster Kahle (brewster@think.com)

• **CORE**: BellCore, OCLC, Cornell, Chem Abstracts, American Chem Society. Exciting project to put all of chemistry online to be used with a small set of chemists. 8 years of almost all chemistry journals are scanned and put on an optical disk jukebox. We have met them in many occasions and they would like to find a way to collaborate with us, but it is unclear how.

Contacts:
Roger H. Thompson (hrothgar@rsch.oclc.org)
Michael E Lesk (lesk@thunder.bellcore.com)
Brewster Kahle (brewster@think.com)

• **Scott and White**: Medical records for researchers. Connection through Dr Argye Hillis got us going in 1989 because the Connection Machine could be used to find similar patients to a new one. This has the potential of helping diagnosis as well as medical research. We have created a large db of free text patient histories (35MB and

200MB). Relevance feedback seems to give very interesting results. Gordon has put in about 1 month on this project.
Contacts: Dr. Argye Hillis, Dr. John Devoracek,
Gordon Linoff (Gordon@think.com).

• OCLC seeking NSF funding to do comparison of CM vs boolean for library records.
Contacts:
Erik Jul (ekj@rsch.oclc.org)
Brewster Kahle (brewster@think.com)

3. Minor collaborations or "too early to tell"

• General Magic

• Software Ventures

• G+N Software

• Assoc of Comp Machinery (ACM) Interested in working on a system to distribute their publications.

• Office of Technology Assessment

• National Library of Medicine. Interested in renovating their Medlar information service, the biggest worldwide in the medical field

• EJV Partners they are a joint venture ($26 million initial capital) of 6 Wall Street brokerage houses (salomon, Morgan, etc). They have set up a Sun network for transaction activities, analysis packages and are looking to WAIS for structuring their information system)

1.g. From many contacts, real prospects

• Telebase is in the last stages of negotiations to buy a machine (8k CM2) to run as the "Homework Helper" service for K-12 children with newspaper, encyclopedia, cliff notes, atlases etc. This will be marketed over Compuserve and other networks starting Q1 93.

Contacts:
Ottavia Bassetti(ottavia@think.com)
Carol Bee-Latty (carol@ Think.com)
Brewster Kahle (brewster@think.com)
Lawrence A Husick (law.husick@AppleLink.Apple.COM)

•**BU/Elsevier/Genesys Partners**. A major electronic publishing experiment through Genesys Partners, as intermediaries, will be done at BU with the distribution of the scientific journals of Elsevier (the second largest publisher of scientific Journals). We would like to work with them on this. Unfortunately the machine that BU has will not be able to support the type of searching that is needed (boolean and therefore the Seeker-Disk product by Harry Morris).
Contacts:
Jim Kollegger Genesys Partners Inc.
Ottavia Bassetti (ottavia@think.com)

•**Bibliotheque de France.** Miterand is spending $2Billion on a new library for France. This will have a significant electronic component (about 3million books scanned, which is more than most libraries have in their complete collections). They will have 3000(?) reading stations (probably NeXT computers) .We are well positioned with the Bibliotheque itself (top decision makers) and have started working (meeting in January 1992) with CAP SESA, a software house that has been chosen last December to be the integrator . This will be a tough sell, but it is a strategic account that would put a CM5 in a very important showcase. Success with this account would have a significant effect in driving French and European sales.
Contacts:
Ottavia Bassetti (ottavia@think.com)
Guy Decaudain (Guyd@think.com)
Beatrice Van Bockstaele (natfred@frmop11.bitnet)

• **Columbia Law School**: putting a large bibliographic database under WAIS. Willem Scholten has done a great deal of development and extension to WAIS to make it appropriate for librarians and law students. They are a major test center for new WestLaw and Lexis (Mead) projects. They now have a WAIS server being used in production by their librarians. Bob Rooney is selling the a CM for this task.

Contacts:
Bob Rooney (rooney@think.com)
Brewster Kahle (brewster@think.com)
Willem Scholten (willem@lawmail.law.columbia.edu)

• **Patent office**. We have had a very encouraging meeting with the patent office 2/92. Their buying schedule is 1996 or so, but they would like to move forward with research on retrieving documents from faulty files (because of Optical character recognition), and with research on automatic classification based on the census work. Further, they would like to use Z39.50 as a standard on their CD-ROM project which is a development project that has an RFP out due on April 15, 1992. This collaboration is very exciting.

Contacts:
Ottavia Bassetti (ottavia@think.com)
Steve Smith (smith@think.com)
Thomas Giammo
Bob Rooney (rooney@think.com)
Art Purcell

• **Dow Jones: DowQuest and IBM replacement**. DowQuest has been less than exciting for Dow Jones because it has not made money for them. We have not been able to help their mainline business of //Text because we have not had a boolean capability. They are using the IBM Stairs product and they want to have a different system in place by July 1993. They are sending out an RFP and we will reply with a CM5 solution. We estimate about a $10M contract will be awarded. We will show them the CM2 Seeker-disk system to show that we are moving concretely in that direction.

Contacts:
Bob Rooney (rooney@think.com)
Brewster Kahle (brewster@think.com)
Tracy Shen (shen@think.com)
Charlie Brady, Dow Jones

•**FAXON** is a $500m company that takes orders from libraries for journals and orders in bulk from many publishers. They want to go into the electronic distribution business, and they are doing it with WAIS. They are hiring 3 people to do the first round of development. They would like a high level meeting with our management to discuss how we can jointly work on this. This meeting has been targeted for March 1992. This is a very exciting prospect for Thinking Machines and WAIS.

Contacts:
Brewster Kahle (brewster@think.com)
Ottavia Bassetti (ottavia@think.com)
Karen Roubicek (roubicek@unifax.faxon.com)

Part 2 Where We are Going: Options and Opportunities for WAIS

WAIS has placed itself at a cross road of two interesting markets for Text Retrieval. The Online Information Services market and the more classical Data Management market.

Text Retrieval in Electronic Publishing: Market Overview

•With the 1991 effort WAIS has established a leadership role for supplying the cutting edge of the $12 billion dollar Online Information services market. If we capitalize on this leadership, we can offer new functionality in a field looking for the next step forward. Doing this will require a creative and flexible approach to keeping our vision and products integrated with other companies and products.

•What we call **"electronic publishing market"** (and that is otherwise defined as "on-line market", "information market place", "electronic information services") is a fast growing arena. Revenues from electronic information services are expected to grow 20% (according to the US industrial outlook of 1991) in the first half of the 90s. Overall the market for online services is estimated at $12 billion for 1992 (source: Information Industry Bulletin).

•Vertical markets in "electronic publishing" are the financial, medical, legal, government and just starting, but growing rapidly, the general interest , educational, and library markets.

•Are the problems big enough for CMs? The answer seems to be yes. Especially if we consider the image database market.
Following are some examples of large Electronic publishers in the US:

Dialog	1 Terabyte of data
Mead Data	400 Gigabytes of data
Dow Jones information services	100 Gigabytes
Patent Office	40 Gigabytes-32 Tbytes with images
UCLA libraries	80 Gigabytes
OCLC catalog	17 Gigabytes

•In this area we have conquered, with WAIS, a leading role in pointing the way. In the past 5 month we have let people come to us and have not done any focused prospecting. We have talked to Mead data, Dialog, Compuserve, OCLC, Faxon, Legislate, and the Washington Post. All of them seem to be ready to change something in their systems and are looking around for new ways to go.

•In the public arena, the groups we have talked to in some interesting terms are the Patent office and the National Library of Medicine. NLM is considering a reorganization of their Medline, one of the fastest growing services in the online market. It's not hard to view more coming up in an area where the government is investing through the NREN program and where all agencies are moving to go online at least for some of their databases.

•The dominating hardware is in the area of mainframes although minicomputers and microcomputers are becoming more important.

•The dominating mainframe software in this type of market remains STAIRS and inhouse versions of it, and more generally boolean search approaches. It seems very hard not to listen to this position of the market: relevance feedback has to be coupled with some boolean research strategy to go ahead in this market. On the other hand, there is generally agreement that boolean will decrease in importance, and offering just boolean is seen as a step backwards.

Text Retrieval in Data Management : a Market Overview

•In the area of large scale information management one trend is starting to define itself: information integration. Imaging, DBMS, Text, maybe voice, are merging together. At the same time standards for integration are starting to emerge.

•For the 90s, (according to a Delphi Consulting Group 1991 Market Survey), the following are the greater demands that will be placed on database management in order of importance.

- Integration
- Multimedia
- Intelligent software
- Media conversion

•The Delphi Consulting Group considers the transition to Text Retrieval and Image databases integration as a **paradigm shift**. (see accompanying materials) In the view of this new paradigm, just as relational databases increased data accessibility over traditional DBMS and 4GLs increased data accessibility over RDBMS, text retrieval systems further increase data accessibility, giving end users access to more information(see picture). Realtime systems further this span by allowing files to be immediately accessible as they come in.

•The dominating hardware is in the area of mainframes although minicomputers and microcomputers will soon take the lead in terms of revenue.

•The overall market for 1990 was in the order of $118 million worldwide for text retrieval, with 35% only for Mini/mainframe in revenues and 25% of the installations, the expectations for 1991 were going up to $300 million in revenues.

Why commit now? Continue to Lead in an Innovative Area as it Grows

•Currently we are leaders in a segment of the Electronic Publishing arena. This gives us advantages in guiding protocols and policy. Even if we are not going to show large short term profits from this area, keeping an active role can be strategically important in the future.

•The WAIS application is very popular which is useful in maintaining our innovative reputation.

•The major decision we have to make are what we want to do within Thinking Machines, and what we want others to do. This is more crucial in this application area since it requires integration of our offerings with other companies, unlike our scientific market. Options here include working with an existing company, forming a joint venture, help start a separate organization, etc.

Why commit now? Push to a Paradigm Shift in DBMS

•Integration of text retrieval and classic database systems (DBMS and RDBMS) products marks a paradigm sift in data management. This can represent a very interesting niche for us to pursue (either as a porting or joint venture) to enhance the functionality of any DBMS software and to take advantage of the R&D already done at Thinking Machines in the area of text retrieval.

•One open question is to decide what to do with Image databases , which is a very attractive area for many University, Defense and Government accounts (for example, the Office of Management and Budget has a 5 year plan that identifies $864 million in five years, starting in 1989 on Imaging projects) and for markets like drug and insurance companies.

Thinking Machine's Options for the Text Market.

We have several options on how we approach this market. It comes in different levels of internal commitment. At several layers we could work with a 3rd party to help create products. We could also be instrumental in creating such a 3rd party.

Level 1: OS Support for Text

Extend the CM5 Operating System to support text applications. These changes, are very similar, if not identical to DBMS needs. In fact, any disk-based applications (unlike scientific applications) need these features:

1) Continue development of CMMD so that this becomes a usable programming environment. I believe we have commitments to do this already, but might need more resources.

2) Allow easy access to individual disks. This is a small subset of "Unix on a Node". Thus files on an individual disk can be opened, read, and closed. This must be made efficient.

3) Support programming on the Scale-array processor with no PN's. We have a full Sun on the Scale-array board, but it is not currently programmable. In many cases, there is no need for processor nodes at all since the scale-array processors are sufficient and much more cost effective for this application. This would further mean to allow the Pathfinder to contain only scale-array boards to make a 32GByte system. A way to do this is to port Unix to the scale-array board.

How hard is this? I believe both 1 and 2 are being done anyway, but this might give an extra push. Option 3 in terms of hardware changes is easy at this point, but gets harder as the hardware gets finalized. The software on option 3 is a couple of man-years. Further study is needed to understand the work required.

Level 2: Text Application Prototyping

Port our text applications to the CM5 and make sure they can run on the OS that we have built. This could go two ways:

1) port the existing Unix server to the CM5 with few modifications.

2) explore ways to make more efficient use of the hardware by spreading out index files on many disks and load balancing on the PN's. Craig Stanfill has starting theoretical studies on how this might be done. 3) do research under government funding.

Option one is just about 6 man-months given that the operating system is functional. This is so short because it leverages the work already done on the Unix search engine. Option 2 is more open ended but would probably be 1-2 man-years. Jonathan Goldman is planning on working on this.

Level 3: Text Application Product Development

Specify an application to build and develop it as a product. There are a couple of options in this case:

1) STAIRS replacement. Take the IBM STAIRS product and make a functionally equivalent version on the CM5. This would target existing information providers. See the market report for details on STAIRS. We already have such a system in test on the CM2.

2) WAIS server bundled on the CM5. This would be usable by any CM5 user for their own data. This means making it easy for people to add new databases and search it with relevance feedback as well as boolean. Integrate into existing environments and be compatible with the WAIS development going on elsewhere.

3) Integrate with a DBMS. A joint Text-DBMS product is the direction that progressive database companies are moving, I am told. This platform would be a powerful and useful one for many uses.

To do this we might work with a 3rd party company to create these products. This would only be possible if the OS were made useful, and probably some amount of prototyping would have to be done by us to make sure the machine is sound and debugged.

How hard are these alternatives?

1) STAIRS alone is not that difficult, but its utility is in question. A guess at effort might be about 3-5 man-years.

2) Porting WAIS is easy since there are no rigorous specifications of what a WAIS server is. Therefore we are free to do either a 2 man-year or a 5 man-year effort in this.

3) This is difficult since it would involve a couple of different groups. I imagine that the DBMS development would be done as a partnership with one of the progressive DBMS manufacturers, while the text work would be done here. I do not know enough to know how this would be done, much less how hard it would be. More study is needed to estimate any of these options.

Level 4: Text Application Marketing

Marketing our text product can be seen as a distinct effort from our scientific marketing since it is to very different communities with different priorities. Some of the communities include: government, law, medicine, businesses, information providers, and publishers. Approaches to marketing:

1) Work with integrators for business and publisher markets.
2) Direct marketing for information providers and government.
3) Work with a 3rd party company that we might help create to handle options 1 and 2.